# Patient Readmission Prediction

## Approach

**Data Preparation & Exploration**: Includes data processing by identifying categorical, numerical and target variables, plotting and analysing distributions and class imbalance.

**Feature Engineering:** Key features were engineered, including

* *stay\_per\_admission* - length of stay relative to previous admissions
* *high\_readmit\_risk* - high readmission risk by prior admissions

**Feature Encoding**

* One-Hot encoding was used for non-ordinal categorical features.

**Model Development and Evaluation**

* **Random Forest Classifier** was used due to ensemble learning benefits, robustness to noise and ability to provide feature importance.
* Compared base model using all features (except multicollinear ones) vs. challenger model with top 5 features based on correlation.

## Key Results & Practical Implications

* Challenger model improves generalisation, smaller train-test performance gap.
* Models are biased towards the negative class which means high readmission risk patients will be missed.
* Challenger model improves test accuracy and ROC AUC (better than random guessing).
* Top influential features:
* *stay\_per\_admission* (0.345038): Higher risk with prolonged stay.
* *age\_group* (0.170): Older patients have higher risk.
* *gender\_Male* (0.059): Slightly more predictive than female suggesting gender bias

## Steps to Address Overfitting & Bias

* Tune model complexity: optimise max\_depth, min\_samples\_leaf.
* Use cross-validation.
* Try alternative models like Logistic Regression or XGBoost.
* Use class\_weight='balanced'.

# Named Entity Recognition (NER) using LLM

## Approach

* Model Selection: Flan-T5 Large from Hugging Face.
* Prompt Engineering: Define extraction needs (diagnosis, treatment, etc.), emphasise medical context, simulate doctor role.
* Application: Use T5Tokenizer on `discharge\_note` to produce `medical\_entities`.

## Key Results & Practical Implications

* 80 non-null results with varying accuracy – potential patient safety risk.
* 120 null results – admin overhead for manual review, delays for patients.

## Steps to Improve NER

* Iteratively refine prompt through engineering such as expanding and refining entity types.
* Fine-tune model on larger, annotated medical datasets.
* Implement robust evaluation with precision, recall, F1-score.

## Risk Considerations

* **Hallucination**: LLMs may generate inaccurate but confident outputs, especially in specialised domains.
* **Entity Ambiguity**: Medical language complexity can lead to misinterpretation.
* **Limitations of General-Purpose Models**: Lack of medical-specific training affects performance.